χ^2 Georgetown Linguistics

Chi Square Tutorial

Prof. Jeff Connor-Linton
Department of Linguistics
Georgetown University
connorlj@gunet.georgetown.edu

Note: For this tutorial to display properly, your browser must support tables and should support superscripts (e.g. 3 squared is '3^2', not '32'). Netscape 2.0 or Mosaic 2.0.1 are recommended.

Topics:

Return to Web Chi Square Calculator

Overview

Chi square is a non-parametric test of statistical significance for bivariate tabular analysis (also known as crossbreaks). Any appropriately performed test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting an hypothesis. Typically, the hypothesis tested with chi square is whether or not two different samples (of people, texts, whatever) are different enough in some characteristic or aspect of their behavior that we can generalize from our samples that the populations from which our samples are drawn are also different in the behavior or characteristic.

A non-parametric test, like chi square, is a rough estimate of confidence; it accepts weaker, less accurate data as input than parametric tests (like t-tests and analysis of variance, for example) and therefore has less status in the pantheon of statistical tests. Nonetheless, its limitations are also its strengths; because chi square is more 'forgiving' in the data it will accept, it can be used in a wide variety of research contexts.

Chi square is used most frequently to test the statistical significance of results reported in bivariate tables, and interpreting bivariate tables is integral to interpreting the results of a chi square test, so we'll take a look at bivariate tabular (crossbreak) analysis.

Bivariate Tabular Analysis

Bivariate tabular (crossbreak) analysis is used when you are trying to summarize the intersections of independent and dependent variables and understand the relationship (if any) between those variables. For example, if we wanted to know if there is any relationship between the biological sex of American undergraduates at a particular university and their footwear preferences, we might select 50 males and 50 females as randomly as possible, and ask them, "On average, do you prefer to wear sandals, sneakers, leather shoes, boots, or something else?" In this example, our independent variable is biological sex. (In experimental research, the independent variable is actively manipulated by the researcher; for example, whether or not a rat gets a food pellet when it pulls on a striped bar. In most sociological research, the independent variable is not actively manipulated in this way, but controlled by sampling for, e.g., males vs. females.) Put another way, the independent variable is the quality or characteristic that you hypothesize helps to predict or explain some other quality or characteristic (the dependent variable). We control the independent variable (and as much else as possible and natural) and elicit and measure the dependent variable to test our hypothesis that there is some relationship between them. Bivariate tabular analysis is good for asking the following kinds of questions:

1. Is there a relationship between any two variables IN THE DATA?

2. How strong is the relationship IN THE DATA?
3. What is the direction and shape of the relationship IN THE DATA?
4. Is the relationship due to some intervening variable(s) IN THE DATA??

To see any patterns or systematic relationship between biological sex of undergraduates at University of X and reported footwear preferences, we could summarize our results in a table like this:

Table 1.a. Male and Female Undergraduate Footwear Preferences

	Sandals	Sneakers	Leather shoes	Boots	Other
Male					
Female					

Depending upon how our 50 male and 50 female subjects responded, we could make a definitive claim about the (reported) footwear preferences of those 100 people.

In constructing bivariate tables, typically values on the independent variable are arrayed on vertical axis, while values on the dependent variable are arrayed on the horizontal axis. This allows us to read 'across' from hypothetically 'causal' values on the independent variable to their 'effects', or values on the dependent variable. How you arrange the values on each axis should be guided "iconically" by your research question/hypothesis. For example, if values on an independent variable were arranged from lowest to highest value on the variable and values on the dependent variable were arranged left to right from lowest to highest, a positive relationship would show up as a rising left to right line. (But remember, association does not equal causation; an observed relationship between two variables is not necessarily causal.)

Each intersection/cell--of a value on the independent variable and a value on the independent variable--reports the result of how many times that combination of values was chosen/observed in the sample being analyzed. (So you can see that crosstabs are structurally most suitable for analyzing relationships between nominal and ordinal variables. Interval and ratio variables will have to first be grouped before they can "fit" into a bivariate table.) Each cell reports, essentially, how many subjects/observations produced that combination of independent and dependent variable values. So, for example, the top left cell of the table above answers the question: "How many male undergraduates at University of X prefer sandals?" (Answer: 6 out of the 50 sampled.)

Table 1.b. Male and Female Undergraduate Footwear Preferences

	Sandals	Sneakers	Leather shoes	Boots	Other
Male	6	17	13	9	5
Female	13	5	7	16	9

Reporting and interpreting crosstabs is most easily done by converting raw frequencies (in each cell) into percentages of each cell within the values/categories of the independent variable. For example, in the Footwear Preferences table above, total each row, then divide each cell by its row total, and multiply that fraction by 100.

Table 1.c. Male and Female Undergraduate Footwear Preferences (Percentages)

	Sandals	Sneakers	Leather shoes	Boots	Other	N
Male	12	34	26	18	10	50
Female	26	10	14	32	18	50

Percentages basically standardize cell frequencies as if there were 100 subjects/observations in each category of the independent variable. This is useful for comparing across values on the independent variable, but that usefulness comes at the price of a generalization--from the actual number of subjects/observations in that column in your data to a hypothetical 100 subjects/observations. If the raw row total was 93, then percentages do little violence to the raw scores; but if the raw total is 9, then the generalization (on no statistical basis, i.e., with no knowledge of sample-population representativeness) is drastic. So you should provide that total N at the end of each row/independent variable category (for replicability and to enable the reader to assess your interpretation of the table's meaning).

With this caveat in mind, you can compare the patterns of distribution of subjects/observations along the dependent variable between the values of the independent variable: e.g., compare male and female undergraduate footwear preference. (For some data, plotting the results on a line graph can also help you interpret the results: i.e., whether there is a positive (/), negative (\), or curvilinear (V, Λ) relationship between the variables.) Table 1.c shows that within our sample, roughly twice as many females preferred sandals and boots as males; and within our sample, about three times as many men preferred sneakers as women and twice as many men preferred leather shoes. We might also infer from the 'Other' category that female students within our sample had a broader range of footwear preferences than did male students.

Generalizing from Samples to Populations

Converting raw observed values or frequencies into percentages does allow us to see more easily patterns in the data, but that is all we can see: what is in the data. Knowing with great certainty the footwear preferences of a particular group of 100 undergraduates at University of X is of limited use to us; we usually want to measure a sample in order to know something about the larger populations from which our samples were drawn. On the basis of raw observed frequencies (or percentages) of a sample's behavior or characteristics, we can make claims about the sample itself, but we cannot generalize to make claims about the population from which we drew our sample, unless we submit our results to a test of statistical significance. A test of statistical significance tells us how confidently we can generalize to a larger (unmeasured) population from a (measured) sample of that population.

How does chi square do this? Basically, the chi square test of statistical significance is a series of mathematical formulas which compare the actual observed frequencies of some phenomenon (in our sample) with the frequencies we would expect if there were no relationship at all between the two variables in the larger (sampled) population. That is, chi square tests our actual results against the null hypothesis and assesses whether the actual results are different enough to overcome a certain probability that they are due to sampling error. In a sense, chi-square is a lot like percentages; it extrapolates a population characteristic (a parameter) from the sampling characteristic (a statistic) similarly to the way percentage standardizes a frequency to a total column N of 100. But chi-square works within the frequencies provided by the sample and does not inflate (or minimize) the column and row totals.

Chi Square Requirements

As mentioned before, chi square is a nonparametric test. It does not require the sample data to be more or less normally distributed (as parametric tests like t-tests do), although it relies on the assumption that the variable is normally distributed in the population from which the sample is drawn.

But chi square, while forgiving, does have some requirements:

1. The sample must be randomly drawn from the population.
2. Data must be reported in raw frequencies (**not percentages**);
3. Measured variables must be independent;
4. Values/categories on independent and dependent variables must be mutually exclusive and exhaustive;
5. Observed frequencies cannot be too small.

1) As with any test of statistical significance, your data must be from a random sample of the population to which you wish to generalize your claims.

2) You should only use chi square when your data are in the form of raw frequency counts of things in two or more mutually exclusive and exhaustive categories. As discussed above, converting raw frequencies into percentages standardizes cell frequencies as if there were 100 subjects/observations in each category of the independent variable for comparability. Part of the chi square mathematical procedure accomplishes this standardizing, so computing the chi square of percentages would amount to standardizing an already standardized measurement.

3) Any observation must fall into only one category or value on each variable. In our footwear example, our data are counts of male versus female undergraduates expressing a preference for five different categories of footwear. Each observation/subject is counted only once, as either male or female (an exhaustive typology of biological sex) and as preferring sandals, sneakers, leather shoes, boots, or other kinds of footwear. For some variables, no 'other' category may be needed, but often 'other' ensures that the variable has been exhaustively categorized. (For some kinds of analysis, you may need to include an "uncodable" category.) In any case, you must include the results for the whole sample.

4) Furthermore, you should use chi square only when observations are independent: i.e., no category or response is dependent upon or influenced by another. (In linguistics, often this rule is fudged a bit. For example, if we have one dependent variable/column for linguistic feature X and another column for number of words spoken or written (where the rows correspond to individual speakers/texts or groups of speakers/texts which are being compared), there is clearly some relation between the frequency of feature X in a text and the number of words in a text, but it is a distant, not immediate dependency.)

5) Chi-square is an approximate test of the probability of getting the frequencies you've actually observed if the null hypothesis were true. It's based on the expectation that within any category, sample frequencies are normally distributed about the expected population value. Since (logically) frequencies cannot be negative, the distribution cannot be normal when expected population values are close to zero--since the sample frequencies cannot be much below the expected frequency while they can be much above it (an asymmetric/non-normal distribution). So, when expected frequencies are large, there is no problem with the assumption of normal distribution, but the smaller the expected frequencies, the less valid are the results of the chi-square test. We'll discuss expected frequencies in greater detail later, but for now remember that expected frequencies are derived from observed frequencies. Therefore, if you have cells in your bivariate table which show very low raw observed frequencies (5 or below), your expected frequencies may also be too low for chi square to be appropriately used. In addition, because some of the mathematical formulas used in chi square use division, no cell in your table can have an observed raw frequency of 0.

The following **minimum frequency thresholds**should be obeyed:

- for a 1 X 2 or 2 X 2 table, expected frequencies in each cell should be at least 5;
- for a 2 X 3 table, expected frequencies should be at least 2;
- for a 2 X 4 or 3 X 3 or larger table, if all expected frequencies but one are at least 5 and if the one small cell is at least 1, chi-square is still a good approximation.

In general, the greater the degrees of freedom (i.e., the more values/categories on the independent and dependent variables), the more lenient the minimum expected frequencies threshold. (We'll discuss degrees of freedom in a moment.)

Collapsing Values

A brief word about collapsing values/categories on a variable is necessary. First, although categories on a variable--especially a dependent variable--may be collapsed, they cannot be excluded from a chi-square analysis. That is, you cannot arbitrarily exclude some subset of your data from your analysis. Second, a decision to collapse categories should be carefully motivated, with consideration for preserving the integrity of the data as it was originally collected. (For example, how could you collapse the footwear preference categories in our example and still preserve the integrity of the original question/data? You can't, since there's no way to know if combining, e.g., boots and leather shoes versus sandals and sneakers is true to your subjects' typology of footwear.) As a rule, you should perform a chi square on the data in its uncollapsed form; if the chi square value achieved is significant, then you may collapse categories to test subsequent refinements of your original hypothesis.

Computing Chi Square

Let's walk through the process by which a chi square value is computed, using Table 1.b. above (renamed 1.d., below).

The first step is to determine our threshold of tolerance for error. That is, what odds are we willing to accept that we are wrong in generalizing from the results in our sample to the population it represents? Are we willing to stake a claim on a 50 percent chance that we're wrong? A 10 percent chance? A five percent chance? 1 percent? The answer depends largely on our research question and the consequences of being wrong. If people's lives depend on our interpretation of our results, we might want to take only 1 chance in 100,000 (or 1,000,000) that we're wrong. But if the stakes are smaller, for example, whether or not two texts use the same frequencies of some linguistic feature (assuming this is not a forensic issue in a capital murder case!), we might accept a greater probability--1 in 100 or even 1 in 20--that our data do not represent the population we're generalizing about. The important thing is to explicitly motivate your threshold before you perform any test of statistical significance, to minimize any temptation for post hoc compromise of scientific standards. For our purposes, we'll set a probability of error thresold of 1 in 20, or p < .05, for our Footwear study.)

The second step is to total all rows and columns:

Table 1.d. Male and Female Undergraduate Footwear Preferences: Observed Frequencies with Row and Column Totals

	Sandals	**Sneakers**	**Leather shoes**	**Boots**	**Other**	**Total**
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

Remember that chi square operates by comparing the actual, or observed, frequencies in each cell in the table to the frequencies we would expect if there were no relationship at all between the two variables in the populations from which the sample is drawn. In other words, chi square compares what actually happened to what hypothetically would have happened if 'all other things were equal' (basically, the null hypothesis). If our actual results are sufficiently different from the predicted null hypothesis results, we can reject the null hypothesis and claim that a statistically significant relationship exists between our variables.

Chi square derives a representation of the null hypothesis--the 'all other things being equal' scenario--in the following way. The expected frequency in each cell is the product of that cell's row total multiplied by that cell's column total, divided by the sum total of all observations. So, to derive the expected frequency of the "Males who prefer Sandals" cell, we multiply the top row total (50) by the first column total (19) and divide that product by the sum total (100): ((50 X 19)/100) = 9.5. The logic of this is that we are deriving the expected frequency of each cell from the union of the total frequencies of the relevant values on each variable (in this case, Male and Sandals), as a proportion of all observed frequencies (across all values of each variable). This calculation is performed to derive the expected frequency of each cell, as shown in Table 1.e below (the computation for each cell is listed below Table 1.e.):

Table 1.e. Male and Female Undergraduate Footwear Preferences: Observed and Expected Frequencies

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male observed	6	17	13	9	5	50
Male expected	9.5	11	10	12.5	7	
Female observed	13	5	7	16	9	50
Female expected	9.5	11	10	12.5	7	
Total	19	22	20	25	14	100

Male/Sandals:	((19 X 50)/100) = 9.5
Male/Sneakers:	((22 X 50)/100) = 11
Male/Leather Shoes:	((20 X 50)/100) = 10
Male/Boots:	((25 X 50)/100) = 12.5
Male/Other:	((14 X 50)/100) = 7
Female/Sandals:	((19 X 50)/100) = 9.5
Female/Sneakers:	((22 X 50)/100) = 11
Female/Leather Shoes:	((20 X 50)/100) = 10
Female/Boots:	((25 X 50)/100) = 12.5
Female/Other:	((14 X 50)/100) = 7

(Notice that because we originally obtained a balanced male/female sample, our male and female expected scores are the same. This usually will not be the case.)

We now have a comparison of the observed results versus the results we would expect if the null hypothesis were true. We can informally analyze this table, comparing observed and expected frequencies in each cell (Males prefer sandals less than expected), across values on the independent variable (Males prefer sneakers more than expected, Females less than expected), or across values on the dependent variable (Females prefer sandals and boots more than expected, but sneakers and shoes less than expected). But so far, the extra computation doesn't really add much more information than interpretation of the results in percentage form. We need some way to measure how different our observed results are from the null hypothesis. Or, to put it another way, we need some way to determine whether we can reject the null hypothesis, and if we can, with what degree of cinfidence that we're not making a mistake in generalizing from our sample results to the larger population.

Logically, we need to measure the size of the difference between the pair of observed and expected frequencies in each cell. More specifically, we calculate the difference between the observed and expected frequency in each cell, square that difference, and then divide that product by the difference itself. The formula can be expressed as:

$((O - E)^2/E)$

Squaring the difference ensures a positive number, so that we end up with an absolute value of differences. If we didn't work with absolute values, the positive and negative differences across the entire table would always add up to 0. (You really understand the logic of chi square if you can figure out why this is true.) Dividing the squared difference by the expected frequency essentially removes the expected frequency from the equation, so that the remaining measures of observed/expected difference are comparable across all cells.

So, for example, the difference between observed and expecetd frequencies for the Male/Sandals preference is calculated as follows:

1. Observed (6) minus Expected (9.5) = Difference (-3.5)

2. Difference (-3.5) squared = 12.25
3. Difference squared (12.25) divided by Expected (9.5) = 1.289

The sum of all products of this calculation on each cell is the total chi square value for the table.

The computation of chi square for each cell is listed below Table 1.f.:

Table 1.f. Male and Female Undergraduate Footwear Preferences: Observed and Expected Frequencies Plus Chi Square

	Sandals	**Sneakers**	**Leather shoes**	**Boots**	**Other**	**Total**
Male observed	6	17	13	9	5	50
Male expected	9.5	11	10	12.5	7	
Female observed	13	5	7	16	9	50
Female expected	9.5	11	10	12.5	7	
Total	19	22	20	25	14	100

Male/Sandals:	$((6 - 9.5)^2/9.5) =$	1.289
Male/Sneakers:	$((17 - 11)^2/11) =$	3.273
Male/Leather Shoes:	$((13 - 10)^2/10) =$	0.900
Male/Boots:	$((9 - 12.5)^2/12.5) =$	0.980
Male/Other:	$((5 - 7)^2/7) =$	0.571
Female/Sandals:	$((13 - 9.5)^2/9.5) =$	1.289
Female/Sneakers:	$((5 - 11)^2/11) =$	3.273
Female/Leather Shoes:	$((7 - 10)^2/10) =$	0.900
Female/Boots:	$((16 - 12.5)^2/12.5) =$	0.980
Female/Other:	$((9 - 7)^2/7) =$	0.571

(Again, because of our balanced male/female sample, our row totals were the same, so the male and female observed-expected frequency differences were identical. This is usually not the case.)

The total chi square value for Table 1 is 14.026.

Interpreting the Chi Square Value

We now need some criterion or yardstick against which to measure the table's chi square value, to tell us whether or not it is significant. What we need to know is the probability of getting a chi square value of a minimum given size even if our variables are not related at all in the larger population from which our sample was drawn. That is, we need to know how much larger than 0 (the absolute chi square value of the null hypothesis) our table's chi square value must be before we can confidently reject the null hypothesis. The probability we seek depends in part on the degrees of freedom of the table from which our chi square value is derived.

Degrees of freedom

Mechanically, a table's degrees of freedom (df) can be expressed by the following formula:

$df = (r-1)(c-1)$

That is, a table's degrees of freedom equals the number of rows in the table minus one multiplied by the number of columns in the table minus one. (For 1 X 2 tables: df = k - 1, where k = number of values/categories on the variable.)

Degrees of freedom is an issue because of the way in which expected values in each cell are computed from the row and column totals of each cell. All but one of the expected values in a given row or column are free to vary (within the total observed--and therefore expected) frequency of that row or column); once the free to vary expected cells are specified, the last one is fixed by virtue of the fact that the expected frequencies must add up to the observed row and column totals (from which they are derived).

Another way to conceive of a table's degrees of freedom is to think of one row and one column in the table as fixed, with the remaining cells free to vary. Consider the following visuals (where X = fixed):

X	X
X	
X	

(r-1)(c-1) = (3-1)(2-1) = 2 X 1 = 2

X	X	X
X		
X		
X		
X		

(r-1)(c-1) = (5-1)(3-1) = 4 X 2 = 8

So, for our Table 1, df = (2-1)(5-1) = 4:

	Sandals	Sneakers	Leather shoes	Boots	Other
Male	X	X	X	X	X
Female	X				

In a statistics book, the sampling distribution of chi square (also know as 'critical values of chi square') is typically listed in an appendix. You read down the column representing your previously chosen probability of error threshold (e.g., p < .05) and across the row representing the degrees of freedom in your table. If your chi square value is larger than the critical value in that cell, your data present a statistically significant relationship between the variables in your table.

Table 1's chi square value of 14.026, with 4 degrees of freedom, handily clears the related critical value of 9.49, so we can reject the null hypothesis and affirm the claim that male and female undergraduates at University of X differ in their (self-reported) footwear preferences.

Statistical significance does not help you to interpret the nature or explanation of that relationship; that must be done by other means (including bivariate tabular analysis and qualitative analysis of the data). But a statistically significant chi square value does denote the degree of confidence you may hold that relationship between variables described in your results is systematic in the larger population and not attributable to random error.

Statistical significance also does not ensure substantive significance. A large enough sample may demonstrate a statistically significant relationship between two variables, but that relationship may be a trivially weak one. Statistical significance means only that the pattern of distribution and relationship between variables which is found in the data from a sample can be confidently generalized to the larger populattion from which the sample was randomly drawn. By itself, it does not ensure that the relationship is theoretically or practically important or even very large.

Measures of Association

While the issue of theoretical or practical importance of a statistically significant result cannot be quantified, the relative magnitude of a statistically significant relationship can be measured. Chi-square allows you to make decisions about whether there is a relationship between two or more variables; if the null hypothesis is rejected, we conclude that there is a statistically significant relationship between the variables. But we frequently want a measure of the strength of that relationship--an index of degree oof correlation, a measure of the degree of association between the variables represented in our table (and data). Luckily, several related measures of association can be derived from a table's chi square value.

For tables larger than 2 X 2 (like our Table 1), a measure called 'Cramer's phi' is derived by the following formula (where N = the total number of observations, and k = the smaller of the number of rows or columns):

Cramer's phi = the square root of (chi-square divided by (N times (k minus 1)))

So, for our Table 1 (a 2 X 5), we would compute Cramer's phi as follows:

1. N(k - 1) = 100 (2-1) = 100
2. chi square/100 = 14.026/100 = 0.14
3. square root of (2) = 0.37

The product is interpreted as a Pearson r (that is, as a correlation coefficient).

(For 2 X 2 tables, a measure called 'phi' is derived by dividing the table's chi square value by N (the total number of observations) and then taking the square root of the product. Phi is also interpreted as a Pearson r.)

A complete account of how to interpret correlation coefficients is unnecessary for present purposes. It will suffice to say that r^2 is a measure called shared variance. Shared variance is the portion of the total behavior (or distribution) of the variables measured in the sample data which is accounted for by the relationship we've already detected with our chi square. For Table 1, $r^2 = 0.137$, so appproximately 14% of the total footwear preference story is explained/predicted by biological sex.

Computing a measure of association like phi or Cramer's phi is rarely done in quantitative linguistic analyses, but it is an important benchmark of just 'how much' of the phenomenon under investigation has been explained. For example, Table 1's Cramer's phi of 0.37 ($r^2 = 0.137$) means that there are one or more variables still undetected which, cumulatively, account for and predict 86% of footwear preferences. This measure, of course, doesn't begin to address the nature of the relation(s) between these variables, which is a crucial part of any adequate explanation or theory.

Return to Web Chi Square Calculator

[Last updated May 18, 1998 by C. Ball]